Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2006

Modelling Pragmatic Aspects of Language Evolution: The Case of the Articulation Bottleneck

Höfler, Stefan

Abstract: Evolutionary linguistics and historical pragmatics can inform each other on a theoretical as well as on a methodological level. This paper introduces a computational model to study the impact of pragmatic factors on language evolution and change. It explores the validity of the frequently cited statement that '[i]t would not be entirely inappropriate to regard languages in their diachronic aspects as gigantic expression-compacting machines' (Langacker, 1977, 107). In particular, I advocate a model of language that includes a coding aspect as well as an inferential aspect, and demonstrate how these two aspects interact with each other through the processes of underspecification and overspecification. The developed model exemplifies how language makes use of lossy compression to adapt to its environment, and how ambiguity functions as a necessary prerequisite for language evolution. By applying the modelled processes to study the origins of language, a new account of protolanguage is put forward to complement the existing approaches to the issue.

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-115032
Conference or Workshop Item
Updated Version

Originally published at:
Höfler, Stefan (2006). Modelling Pragmatic Aspects of Language Evolution: The Case of the Articulation Bottleneck. In: Linguistics Association of Great Britain Annual Meeting (LAGB 2006), Newcastle upon Tyne, 30 August 2006 - 2 September 2006, s.n..

Modelling Pragmatic Aspects of Language Evolution: The Case of the Articulation Bottleneck *

STEFAN HOEFLER

Language Evolution and Computation Research Unit
Linguistics and English Language, University of Edinburgh
stefan@ling.ed.ac.uk

Evolutionary linguistics and historical pragmatics can inform each other on a theoretical as well as on a methodological level. This paper introduces a computational model to study the impact of pragmatic factors on language evolution and change. It explores the validity of the frequently cited statement that '[i]t would not be entirely inappropriate to regard languages in their diachronic aspects as gigantic expression-compacting machines' (Langacker, 1977, 107). In particular, I advocate a model of language that includes a coding aspect as well as an inferential aspect, and demonstrate how these two aspects interact with each other through the processes of underspecification and overspecification. The developed model exemplifies how language makes use of lossy compression to adapt to its environment, and how ambiguity functions as a necessary prerequisite for language evolution. By applying the modelled processes to study the origins of language, a new account of protolanguage is put forward to complement the existing approaches to the issue.

1. Introduction

Recent years have seen the emergence of two new interdisciplinary linguistic subfields: evolutionary linguistics and historical pragmatics. While evolutionary linguistics is concerned with the origins and evolution of language, historical pragmatics, or at least one major branch of it, focuses on the impact of pragmatic factors on language change. It is one aim of this paper to show how the introduction of pragmatic factors can also have a substantial effect on models of language evolution.

However, it is not only historical pragmatics that can inform evolutionary linguistics. The exchange can be seen as working both ways. Both, historical pragmatics as well as evolutionary linguistics share a data problem. The object of study of evolutionary linguistics, the origins of language, is simply too far remote in the past to have provided any record or evidence (Cangelosi & Parisi, 2001).

*This paper was presented at the *Linguistics Association of Great Britain Annual Meeting (LAGB 2006)*, Newcastle upon Tyne, 30 August – 2 September 2006.

Historical pragmatics on the other side, like any branch of pragmatics, is primarily interested in processes and phenomena that are centred in spoken language. But spoken language of the historical past cannot be investigated through direct observation, since our present-day recording techniques were not available in earlier times (Jacobs & Jucker, 1995). Evolutionary linguistics has found ways to overcome its data problem to a certain extent by developing computational models (Cangelosi & Parisi, 2001). This paper exemplifies how such models can equally be used to formalise and test theories assigned to historical pragmatics, concerned with the impact of pragmatic factors on language change.

In particular, I will evaluate the frequently cited statement by Langacker (1977, 107) that '[i]t would not be entirely inappropriate to regard languages in their diachronic aspects as gigantic expression-compacting machines' in the light of pragmatic factors influencing language evolution and change.

In the remainder of this paper, I will introduce the articulation bottleneck as a constraint on language evolution and explain why this property of language necessitates an extension of the model of language typically adopted in evolutionary linguistics by the introduction of context (sections 2 and 3). I will then describe two central processes in such a model: underspecification (section 4) and overspecification (section 5), and show how they allow language to adapt to its environment. Special attention will be given to the evolutionary role of ambiguity, both in the original and in the extended model of language (section 6). The paper concludes with an application of the developed processes to questions of the emergence of protolanguage (section 7) and a summary of the findings and the main claims that can be derived from them (section 8).

2. The articulation bottleneck

Numerous studies in evolutionary linguistics are concerned with bottlenecks in language. More specifically, they typically look at the evolutionary dynamics arising from one particular bottleneck, namely the *learning bottleneck* (e.g. Hurford, 2002; Smith, Kirby, & Brighton, 2003). The learning bottleneck is a description of the fact that language has to be learned repeatedly: each generation of speakers learns their language from the previous generation. In order to persist, language therefore has to be learnable—and has evolved to do so, as some studies suggest (Brighton, Kirby, & Smith, 2005)—in order to pass the learning bottleneck.

In this article, however, I will focus on another, equally fundamental bottleneck: the *articulation bottleneck*. The articulation bottleneck is the property of language that it transmits messages through a serial channel, speech. Levinson (1995, 95f.) identifies articulation as a 'relatively slow and inefficient process, which acts as a bottleneck in the entire communicative procedure'. To provide evidence, he mentions that we can think faster than we can speak, that we can easily understand pitch-corrected speech at double speed, and that we can scan a printed page far faster than it can be read aloud. It is the process of articulation that slows down the transmission system of human verbal communication. This fact is reflected in various principles and heuristics introduced by pragmatists, ranging from the well-known conversational maxims 'Do not make you contribution more informative than is required' and 'Be brief' (Grice, 1975) to the assertion that '[i]t will pay to say little and infer much' (Levinson, 1995, 96). However, the articulation bottleneck is first and foremost a physical and information-theoretic condition of language: in order to be transmitted via speech, elements of meaning have to be transformed—i.e. *compressed*—into information-bearing units of form.

Two caveats seem to be required at this point. First, the articulation bottleneck must not be confused with another concept of a similar name, the 'ease of articulation'. The ease of articulation is a physiological property that may but need not lead to a compression of the signal, and therefore does not stand in a *direct* relation to the articulation bottleneck. The second distinction we have to be aware of is that of *signal simplicity* as opposed to *code simplicity*. The articulation bottleneck conditions the compression of the speech *signal*, rather than that of the language *code* or its grammar. The evolutionary role of the latter has been discussed elsewhere (e.g. Brighton, 2003), and the related notions of representational economy and processing economy are particularly popular in the generativist literature. Langacker (1977) provides an extensive discussion of the distinction between signal simplicity and code simplicity.

3. Compression and the code model of language

In the previous section, we have seen that the articulation bottleneck is the condition that requires language to transform elements of meaning into elements of form to transmit them via speech. This is a process of *data compression*. Information theory distinguishes two basic types of compression (Sayood, 2006). *Lossless compression* tries to minimise *statistical* redundancy in a signal. It can be realised, for instance, by using shorter words for more frequent meanings. Within linguistic phenomena, Zipf's law (Zipf, 1936) and anaphora are examples that fall under the category of lossless compression.

A far greater compression rate, however, can be achieved by so-called *lossy compression*. Lossy compression seeks to minimise *perceptual* redundancy. This can be done by omitting information, or elements of meaning, which are either imperceptible or inferable from the context.[a] Lossy compression is thus relative to *context*. To discover if lossy compression is at work in language, we must therefore include the notion of context in our model of language.

Information theory—and to a large extent evolutionary linguistics too—typically adheres to a mere code model of language. Language is conceived as a

[a]In the technical domain, lossy compression is applied, for instance, in TV signals, where colour information that cannot be perceived by the human eye is not transmitted, or similarly in the file formats of JPEG and MPEG (Sayood, 2006).

code, and verbal communication accordingly as encoding and decoding of meaning. One problem with the code model of language is that of ambiguity (Hoefler, 2006). If we adopt a conception of language as a mere code, then ambiguity would seem to be a dysfunctional feature. In an optimal code, one must assume, one signal corresponds to exactly one meaning. Furthermore, the fact that there is such an apparent imperfection in language poses another problem, namely an evolutionary one: why has an apparently dysfunctional feature like ambiguity emerged and managed to persist in language? Is it an adaptation of language to some condition we have not been aware of so far, or is it a side-effect of some other property of language? And given our previous considerations, we must also ask if the problem actually persists if we go beyond the code model of language.

It is thus for two reasons that we want to transcend the mere code model of language. First, we want to introduce the notion of context to see if lossy compression can be observed to be at work in language. Second, we hope to overcome the evolutionary problem that ambiguity poses.

Proponents of the code model of language have frequently pointed out that language is essentially a set of mappings from meaning to form (e.g. Pinker & Bloom, 1990). And also those who hold the code model for insufficient and incomplete do not question that language does contain a coding part (Sperber & Wilson, 1995, 9). It is thus safe to define the first two building blocks of our model as an inventory of elements of meaning and an inventory of elements of form. In language, elements of meanings comprise lexical meanings (e.g. TREE, TOP, ...) and grammatical functions (e.g. PLURAL, PAST, ...). In our model, we represent them as lower-case letters (*a, b, c, ...*). Elements of form, on the other side, correspond to the information-bearing units of form such as phonemes, tone, or word order. In our model, we implement just two abstract units of form, namely the binary numbers *0* and *1*. A language user's grammar or code will therefore consist of a set of constructions, which are mappings from a number of elements of meaning onto a number of elements of form.

In a code model of language, constructions are used by the speaker to encode a meaning he or she wants to communicate in a signal. And they are likewise used by the hearer to decode the transmitted signal and retrieve the intended meaning. Fig. 1 illustrates this process.

4. Underspecification

Having established the basic coding part of our model of language, we can now move on to including context. Context can be described as a *cognitive* environment *shared* by speaker and hearer consisting of a set of facts which are *perceptible* or *inferable*. Sperber and Wilson (1995, 15f.) point out that context contains more than just information about the immediate physical environment and information about the immediately preceding utterances. Our conception of context must include, among other things, notions such as world knowledge (including

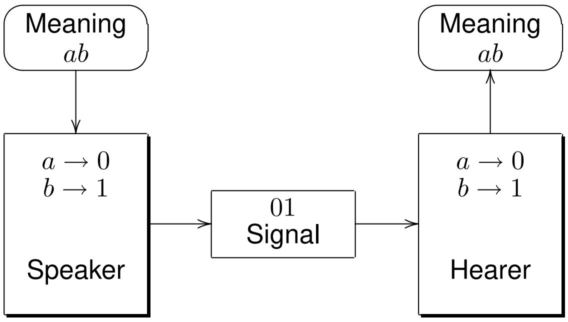

Figure 1. Communication in a code model of language. Both, the speaker's and the hearer's language code contain the two constructions $a \longrightarrow 0$ and $b \longrightarrow 1$. Using these constructions, the speaker *encodes* the intended meaning ab and produces the corresponding signal 01. The hearer *decodes* the transmitted signal and thus recovers the intended meaning ab.

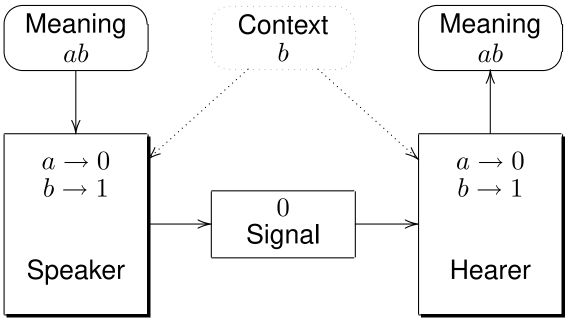

Figure 2. Underspecification. The speaker wants to communicate meaning ab. Since b is in the context, either directly perceptible or inferable given a, the intended meaning ab can be *underspecified*, and thus compressed lossily, by only encoding a and omitting b. The speaker rightly assumes that the hearer will be able to restore the intended meaning by decoding the signal for a and inferring b from the context.

linguistic knowledge), general cultural assumptions, and believes about the state of mind of the speaker or hearer. In our model, context accordingly consists of a set of elements of meanings which are understood as being either perceptible or inferable. If a speaker wants to express a meaning ab, he or she can now *underspecify* their signal if b is either directly perceptible in the context or inferable from it given a. In either way, the speaker can achieve a minimisation of the signal by confining himself to encoding a and assuming that the hearer will be capable of inferring b from context. This process of underspecification is illustrated in Fig 2.

Underspecification thus yields compressed signals by omitting information relative to context. We can therefore identify it as the way language applies lossy compression. Underspecification underlies a vast number of linguistic phenomena linked to the elliptical nature of speech or the vagueness of language.

If the same type of underspecification occurs frequently in similar contexts, a new construction $ab \rightarrow 0$ is added to the language. The inferred part of the intended meaning becomes codified and can now be directly accessed in the code. This process is generally known as *conventionalisation*. In a computer simulation, the newly codified construction can be added to the grammar of the speaker or of the hearer, depending on what class of models of language one favours. If one intends to emphasise the role of learning, one might add the construction as a newly acquired part to the language of the hearer. Linguists who prefer usage-based models, on the other side, will identify the same process as entrenchment or routinisation and locate the addition of the new construction in the speaker. For the purpose of describing language change, the latter type of models can in fact completely ignore the hearer and just focus on the innovations created by the speaker when producing signals. Whichever model one prefers, as far as the evolution and changing of language is concerned, the result is the same. Both variants are thus equivalent for our purposes, and conventionalisation will be represented in the speaker's as well as in the hearer's grammar in this paper.

We have seen that underspecification is language applying lossy compression. We can now state that conventionalisation is the way language adapts to its environment. This environment, however, must not be confused with the context of an individual utterance. The environment of language is rather a complex constellation of parameters consisting of the frequencies with which individual elements of meaning occur in contexts and the frequencies with which they occur in meanings speakers want to express. Furthermore, these frequency distributions are themselves subject to change. Language constantly adapts to a *dynamic* environment and will therefore not reach a state where it remains static.

As a summary of our findings so far, we can state that Langacker accurately describes the interaction of language code and context when he claims that '[i]t would not be entirely inappropriate to regard languages in their diachronic aspects as gigantic expression-compacting machines'. It is underspecification and subsequent conventionalisation that perform this function. Our findings are corroborated by studies in the expression minimisation and grammaticalisation of individual languages (e.g. Li, 2002).

5. Overspecification

It has been shown that compression without counterbalance reduces the expressivity of a language to its ultimate collapse (Brighton & Kirby, 2001), and historical linguists have pointed out that the pressure for simplicity or economy is countervailed by a pressure for *transparency* or *expressivity* (Langacker, 1977).

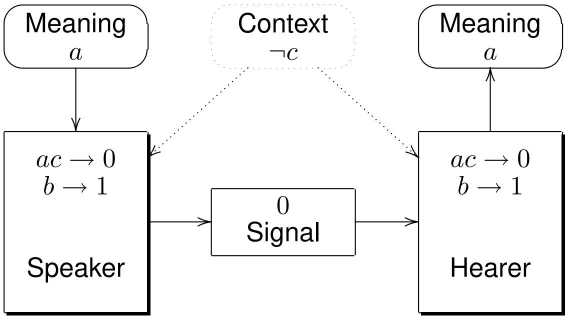

Figure 3. Overspecification. The speaker intends to express the meaning a but has only access to constructions encoding ac and b. Given that it can be inferred from context that c is not the case, the intended meaning can be *overspecified* by transmitting the signal encoding ac. The hearer will decode the transmitted signal as meaning ac but, since it can be inferred from the context that c cannot be the case, recover the right meaning a.

In our model, this second driving-force can be observed in a situation where a speaker intends to express a meaning a but has only access to constructions for ac and b, as illustrated in Fig. 3. The original meaning a can be encoded in an *overspecified* way, using the signal for ac instead, if it can be inferred from the context that c is *not* the case. The speaker assumes that the hearer will decode the transmitted signal as meaning ac but will also infer from the context that c does not hold and therefore be capable of recovering the intended meaning a. Like underspecification, overspecification is a general process that underlies a large number of linguistic phenomena. Maybe the most prominent among them occurs when a lexeme is used in a novel metaphorical way.

In fact, the element c need not even be negated in the context. It will suffice if it is just irrelevant. The speaker and the hearer will ignore it. The conventionalisation of this type of overspecification is known as bleaching, where a morpheme looses parts of its meaning that have become unimportant.

6. Ambiguity and contextual plasticity

In Fig. 3, the introduction of the newly conventionalised construction $a \rightarrow 0$ will obviously lead to an ambiguous situation, where 0 is mapped onto ac as well as onto the more general meaning a. One might argue that one of the two constructions will end up to be used less frequently and eventually disappear from that language, thus rendering an unambiguous situation again. However, linguistic evidence shows that this is not always the case. In fact, layering, the state where an older and a newer meaning for the same form coexist, is a frequently observed phenomenon (e.g. Hopper & Traugott, 1993, 124–126). Fig. 4 shows

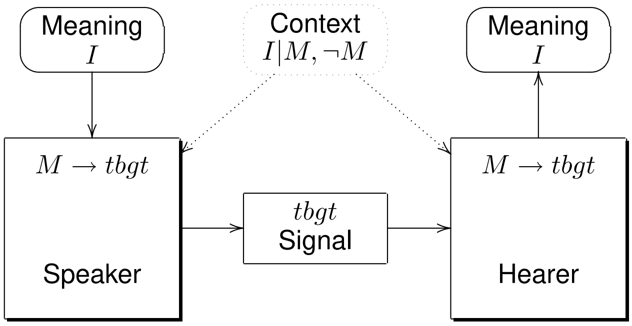

Figure 4. Under- and overspecification combined. In the initial code, TO BE GOING TO ($tbgt$) denotes physical movement (M) only. However, it can be inferred from the context that physical movement implies intention ($I|M$), and that physical movement cannot be part of the communicated message ($\neg M$). Trying to express intention (I), the speaker overspecifies by transmitting the signal for physical movement, assuming that the hearer will be able to infer from the context that physical movement cannot be what the speaker is communicating. At the same time, the speaker underspecifies by assuming that the hearer can infer the intention part from the context once he has decoded the signal for physical movement. Conventionalisation will lead to an (ambiguous) situation of layering, where $tbgt$ is mapped onto M as well as onto I.

how the processes of under- and overspecification can have led to such a situation of layering in the English expression TO BE GOING TO.

However, the crucial point is that conventionalisation can only take place if language allows for ambiguity. If our model prevented ambiguous states, novel interpretations of a signal could not be codified. The possibility for ambiguity is a necessary prerequisite for conventionalisation. Without it, there would thus be no evolution of the coding part of language in the way we have just described in a model of language that consists of a coding aspect and an inferential aspect.

We have also seen that the two aspects of language are not isolated but interact with each other through under- and overspecification, which together constitute a property of language I propose to call *contextual plasticity*. This concept is related, but not identical, to notions such as ‘loose talk’ (Sperber & Wilson, 1995, 233ff.) or ‘pragmatic polysemy’ (Traugott & Dasher, 2005, 11ff.). Contextual plasticity works by shifting elements of meaning into the code through conventionalisation of underspecification, or by going the opposite way when conventionalising overspecification (Fig. 5).

7. From contextual plasticity to protolanguage

Can we apply the described processes to studying the origins of language or its very early stages, which are usually called protolanguage? It seems intuitive to

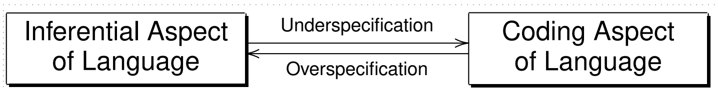

Figure 5. Contextual plasticity. Language exhibits contextual plasticity in the fact that it consists of a coding aspect and an inferential aspect, and that these two aspects interact with each other via under- and overspecification. Conventionalisation of underspecification brings about a shift of elements of meaning into the coding part (codification), whereas conventionalisation of overspecification has the opposite effect.

propose an initial, pre-language situation that only knows the inferential aspect but does not include a code yet. We have seen above that the mechanism we need in order to shift elements of meaning from the inferential aspect of language into the code is underspecification. We must therefore assume a situation of *maximal* underspecification. Such a situation will be one in which the *whole* intended meaning can be inferred from the context. All a speaker will then have to do to communicate that meaning is to claim the hearer's attention, assuming that, once he or she will have managed to have the hearer's attention and once they will therefore share a cognitive environment, the hearer will be able to infer from the context what the hearer intends to communicate. This is in line with Sperber and Wilson (1995) who sate that communicating is claiming an individual's attention. Fig. 6(a) represents the described situation in our model.

At this pre-language stage, speaker and hearer do not have any codified language yet but rely on an inventory of simple signals to claim each other's attention that are not mapped onto any meaning. Likewise, it appears to be rather counter-intuitive to speak of conventionalisation at this early stage of language evolution. But it can be assumed that language users will be capable of memorising the situations in which particular attention claimers occurred and therefore of associating communicated meanings with individual signals. Storing such associations, however, is equivalent to having a code in that overspecification can be applied to them just as it would be applied to a code. Under the right conditions—i.e. if all elements of such a stored meaning the speaker does not want to express are negated in the context—an overspecified signal can now be transmitted and conventionalisation can apply, as illustrated in Fig. 6(b). Our computer simulations confirm that the evolution of the language code can indeed take off the ground if these conditions are met.

As opposed to the *holistic* approach to protolanguage (Wray, 1998), which assumes an initial state of complex signals denoting complex meanings, and the *synthetic* approach (Bickerton, 1990, 2003), beginning from simple signals standing for simple meanings, our approach—we may call it *inferential*—proposes an initial situation in which simple signals (attention claimers that are not mapped

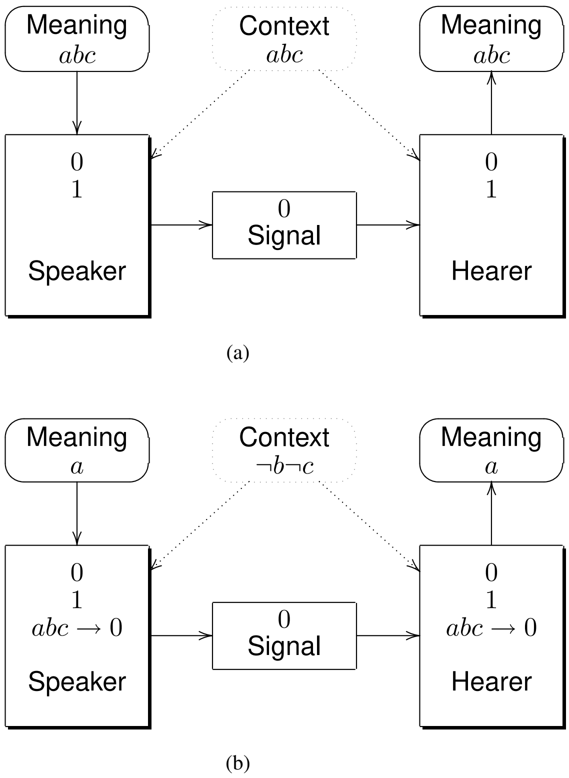

Figure 6. Protolanguage. (a) Maximal underspecification. Speaker and hearer do not have a codified language yet but are equipped with an inventory of attention claimers which are not mapped onto any meaning (0, 1). Communication happens by the speaker claiming the hearer's attention and the hearer inferring what the speaker intends to communicate from context. (b) The situation in which the previous attention claimer occurred has been memorised, and the signal associated with the meanings communicated in that situation. Overspecification is applied to the stored association to express a novel, less specific meaning.

onto any meaning) initialise the inference of maximally underspecified rich meanings and are subsequently associated with them. Language then moves easily to signals denoting simpler, more general meanings through the process of overspecification.

8. Conclusion

I have begun this paper by discussing bottlenecks in language evolution. Rather than looking at the extensively studied learning bottleneck, this paper focuses on the evolutionary dynamics arising from the articulation bottleneck, the property of language that it transmits thoughts via a serial channel, speech. Considering notions of compression as described in information theory, we have seen that lossy compression, which achieves higher compression rates than lossless compression, is relative to context. We have therefore extended our model of language from a mere code model to one that consists of a coding part as well as an inferential part. I have illustrated how these two aspects of language interact with each other through contextual plasticity, i.e. through the processes of under- and overspecification. While ambiguity appears the be a dysfunctional property in a mere code model of language, it is a necessary prerequisite for the evolution of language in our extended model. Finally, I have sketched how contextual plasticity can be applied to a new understanding of the processes involved in the emergence of protolanguage and the subsequent evolution of a linguistic code. To conclude, I propose that

- *if* we adopt a model of language that includes a coding aspect as well as an inferential aspect,
- *and if* we allow these two aspects to interact through contextual plasticity,
- *then* we are able to observe how language adapts to its environment through lossy compression,
- *and* we can dissolve the evolutionary problem that ambiguity poses in a mere code model of language.

Acknowledgements

Stefan Hoefler was supported by Cogito Foundation Research Scholarship 101/04.

References

Bickerton, D. (1990). *Language and Species.* University of Chicago Press.

Bickerton, D. (2003). Symbol and structure: a comprehensive framework for language evolution. In M. Christiansen & S. Kirby (Eds.), *Language Evolution: The States of the Art.* Oxford University Press.

Brighton, H. (2003). *Simplicity as a driving force in linguistic evolution.* Unpublished doctoral dissertation, The University of Edinburgh, Edinburgh.

Brighton, H., & Kirby, S. (2001). *Meaning space structure determines the stability of culturally evolved compositional language* (Tech. Rep.). Language Evolution and Computation Research Unit, Department of Theoretical and Applied Linguistics, The University of Edinburgh.

Brighton, H., Kirby, S., & Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), *Language Origins: Perspectives on Evolution.* Oxford: Oxford University Press.

Cangelosi, A., & Parisi, D. (2001). Computer simulation: A new scientific approach to the study of language evolution. In A. Cangelosi & D. Parisi (Eds.), *Simulating the Evolution of Language* (pp. 3–28). London: Springer.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics: Speech Acts* (Vol. 3, pp. 41–58). New York: Academic Press.

Hoefler, S. (2006). Why has ambiguous syntax emerged? In A. Cangelosi, A. D. M. Smith, & K. Smith (Eds.), *The Evolution of Language: Proceedings of the 6th International Conference on the Evolution of Language (EVOLANG6)* (pp. 123–130). World Scientific.

Hopper, P. J., & Traugott, E. C. (1993). *Grammaticalization.* Cambridge: Cambridge University Press.

Hurford, J. R. (2002). Expression/induction models of language evolution: dimensions and issues. In T. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models.* Cambridge University Press.

Jacobs, A., & Jucker, A. H. (1995). The historical perspective in pragmatics. In A. H. Jucker (Ed.), *Historical Pragmatics* (pp. 3–33). Amserdam, Philadelphia: John Benjamins.

Langacker, R. W. (1977). Syntactic reanalysis. In C. N. Li (Ed.), *Mechanisms of Syntactic Change* (pp. 57–139). Austin: University of Texas Press.

Levinson, S. C. (1995). Three levels of meaning. In F. R. Palmer (Ed.), *Grammar and Meaning* (pp. 90–115). Cambridge: Cambridge University Press.

Li, I. C. (2002). Minimization, conversational inference, and grammaticalization in Taiwanese Southern Min. *Concentric: Studies in Linguistics*, *28*(2), 117–148.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*(4), 707–784.

Sayood, K. (2006). *Introduction to Data Compression* (3rd ed.). San Francisco: Morgan Kaufmann.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated Learning: a framework for the emergence of language. *Artificial Life*, *9*(4), 371-386.

Sperber, D., & Wilson, D. (1995). *Relevance: Cognition and communication* (2nd ed.). Oxford: Blackwell.

Traugott, E. C., & Dasher, R. B. (2005). *Regularity in Semantic Change.* Cambridge: Cambridge University Press.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication*, *18*(1), 47-67.

Zipf, G. K. (1936). *The Psycho-Biology of Language: An Introduction to Dynamic Phonology.* London: Routledge.